The Linear Representation Hypothesis and the Geometry of Large Language Models

Year: 2023 | *Citations: 303* | *Authors: Kiho Park, Yo Joong Choe, Victor Veitch*

Abstract

Informally, the 'linear representation hypothesis' is the idea that high-level concepts are represented linearly as directions in some representation space. In this paper, we address two closely related questions: What does"linear representation"actually mean? And, how do we make sense of geometric notions (e.g., cosine similarity or projection) in the representation space? To answer these, we use the language of counterfactuals to give two formalizations of"linear representation", one in the output (word) representation space, and one in the input (sentence) space. We then prove these connect to linear probing and model steering, respectively. To make sense of geometric notions, we use the formalization to identify a particular (non-Euclidean) inner product that respects language structure in a sense we make precise. Using this causal inner product, we show how to unify all notions of linear representation. In particular, this allows the construction of probes and steering vectors using counterfactual pairs. Experiments with LLaMA-2 demonstrate the existence of linear representations of concepts, the connection to interpretation and control, and the fundamental role of the choice of inner product.